**REPORT**

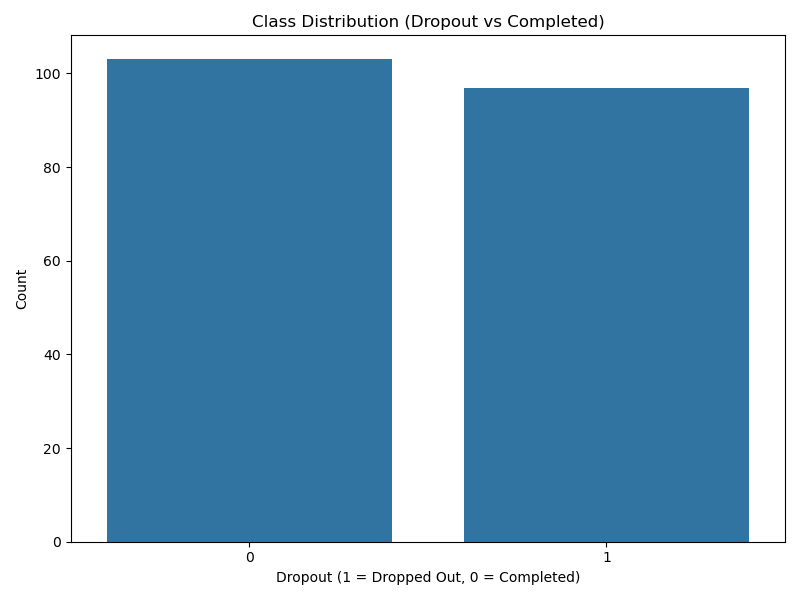
**Data Insights**

• Dataset includes features like **age**, **gender**, **course\_type**, **session\_count**, **quiz\_attempts**, etc.

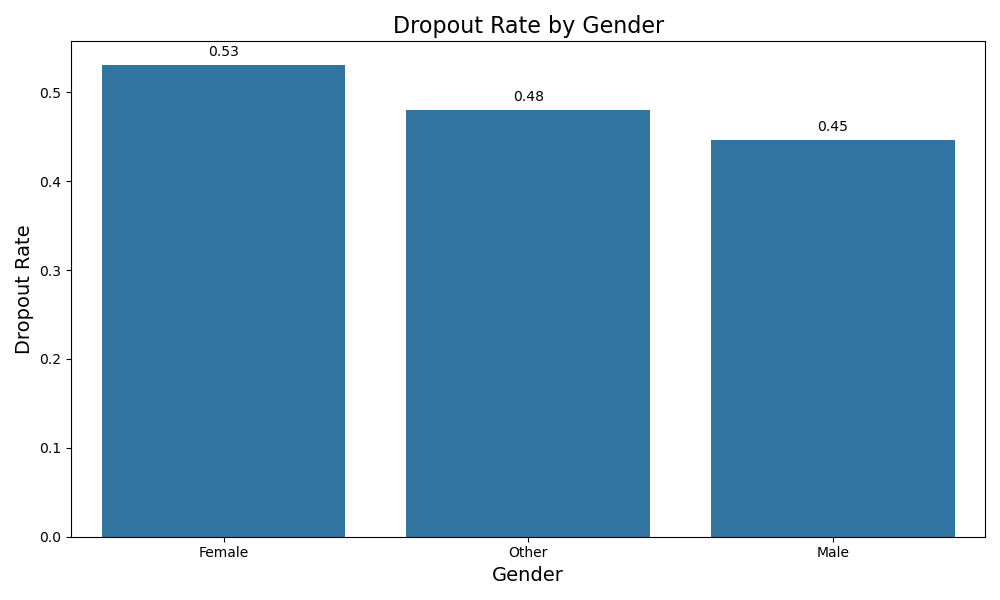
• No missing values; median/mode imputation was included for robustness.

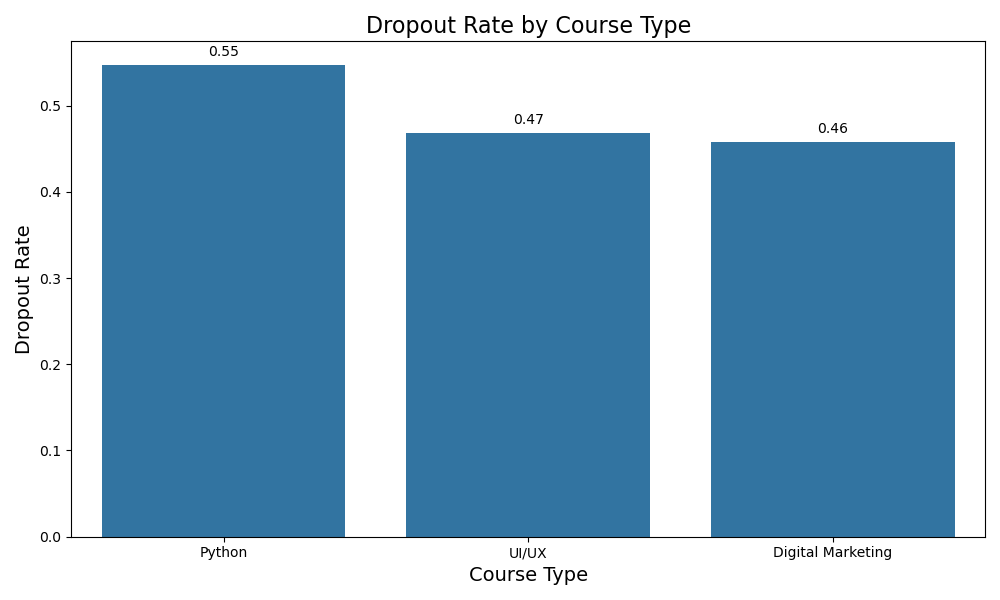
• Visuals suggest:

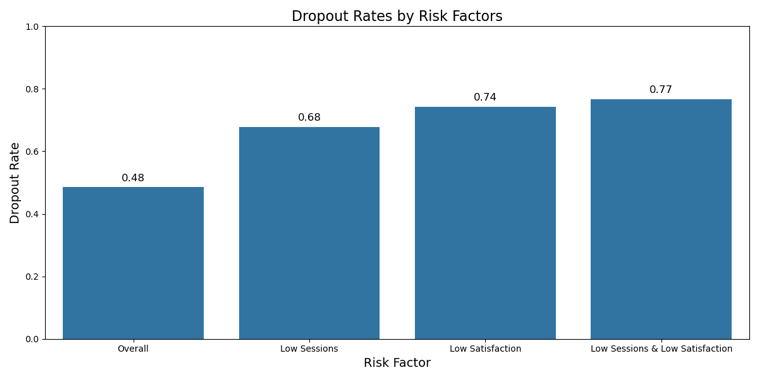
• Imbalanced class distribution (dropout ≫ completion)

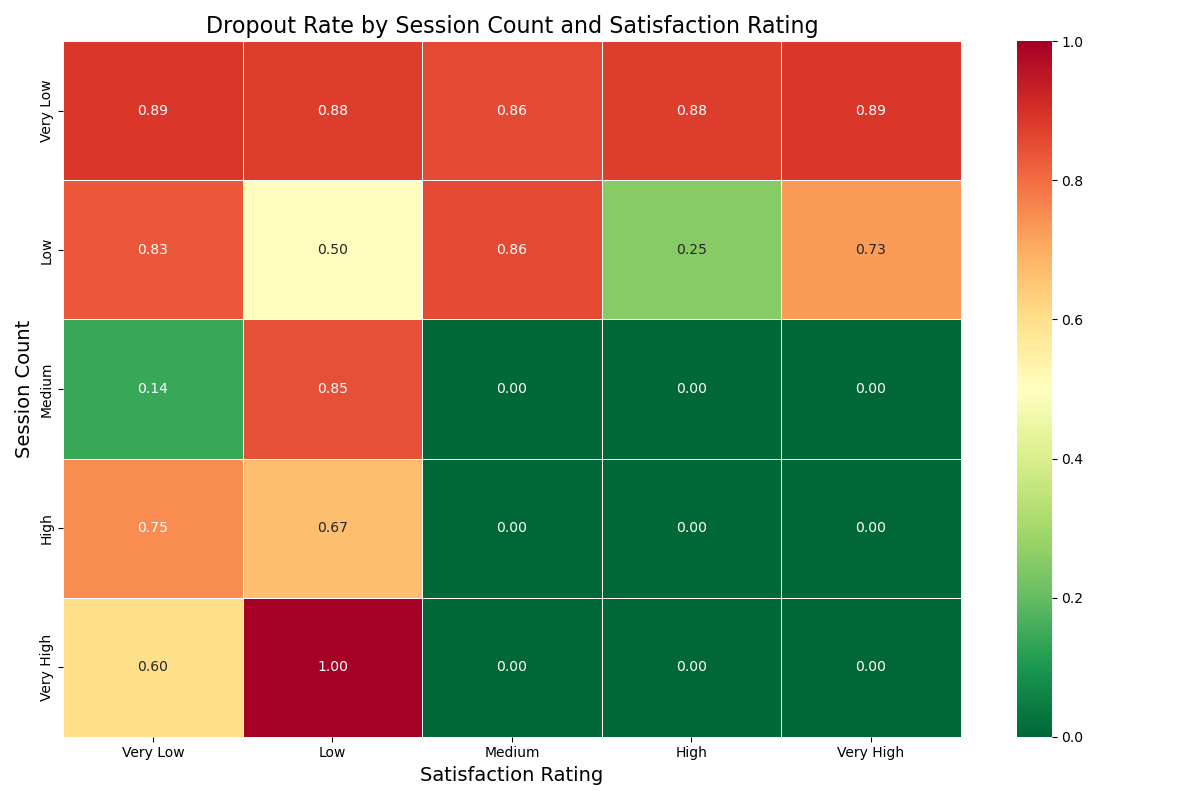


• The below are the dropout rates across categories

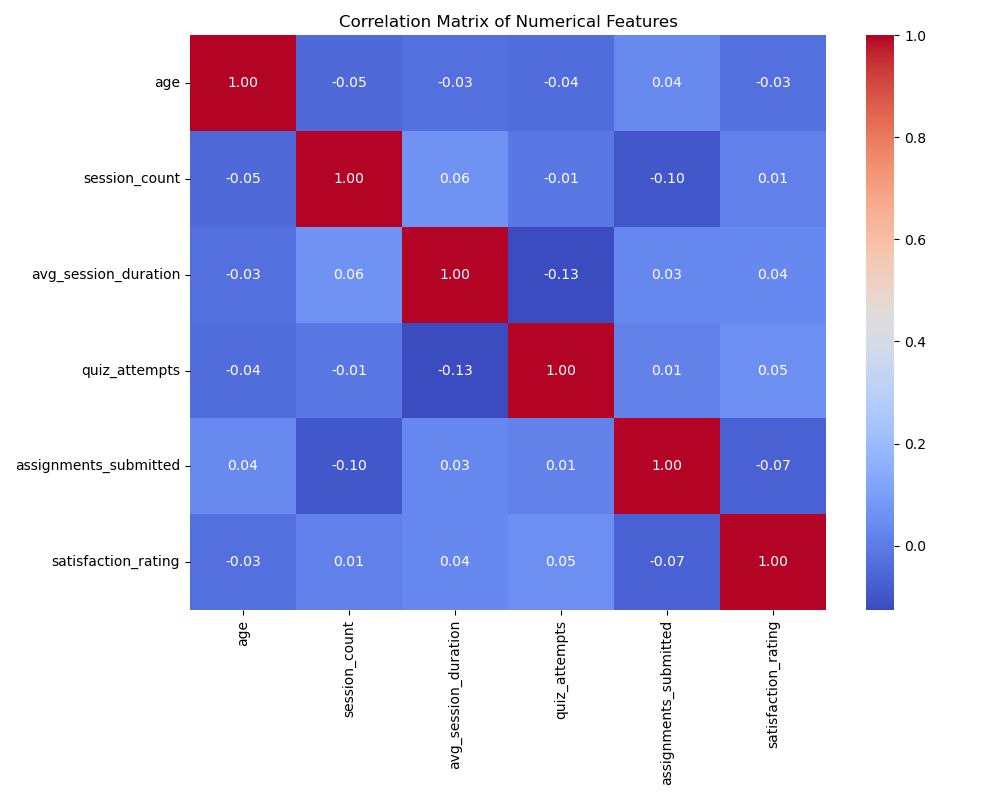








• Strong correlation between session\_count, avg\_session\_duration, and engagement metrics

****

**Modelling Approach**

**Preprocessing**

• **Categorical Encoding:**

Categorical features like gender and course\_type were encoded using **OneHotEncoding**, ensuring non-linear models like Random Forest and XGBoost could learn from them without introducing ordinal bias.

• **Feature Scaling:**

Continuous variables such as session\_count, avg\_session\_duration, and quiz\_attempts were normalized using **StandardScaler**. This ensured stable convergence in linear models and enhanced comparability across features.

• **Train-Test Split:**

A stratified train-test split (typically 80-20) preserved the original class distribution to ensure fair evaluation.

• **Pipeline Integration:**

A unified preprocessing pipeline was built using ColumnTransformer and Pipeline to avoid data leakage and simplify deployment.

**Target Variable**

• The binary target variable dropout indicated whether a student discontinued the course (1) or successfully completed it (0).

• Class imbalance was addressed by model tuning, use of evaluation metrics like **F1-score** and **AUC**, and class-weight adjustments where applicable.

**Models Trained**

1. **Logistic Regression**

• Served as a baseline for linear separability.

• Provided interpretability through coefficient weights.

• Limitation: struggled with complex non-linear relationships.

2. **Random Forest Classifier**

• Ensemble method leveraging bagging and decision trees.

• Naturally handled non-linear patterns and interactions.

• Provided feature importance insights.

• Tuned via n\_estimators, max\_depth, and min\_samples\_split.

3. **XGBoost (Extreme Gradient Boosting)**

• High-performance gradient boosting model.

• Captured subtle feature interactions with excellent generalization.

• Outperformed all other models in precision, recall, and F1.

• Tuned with early stopping, regularization (lambda, alpha), and learning rate control.

**Evaluation Results**

• **Logistic Regression:**

Accuracy: 0.81 | Precision: 0.78 | Recall: 0.70 | F1: 0.74

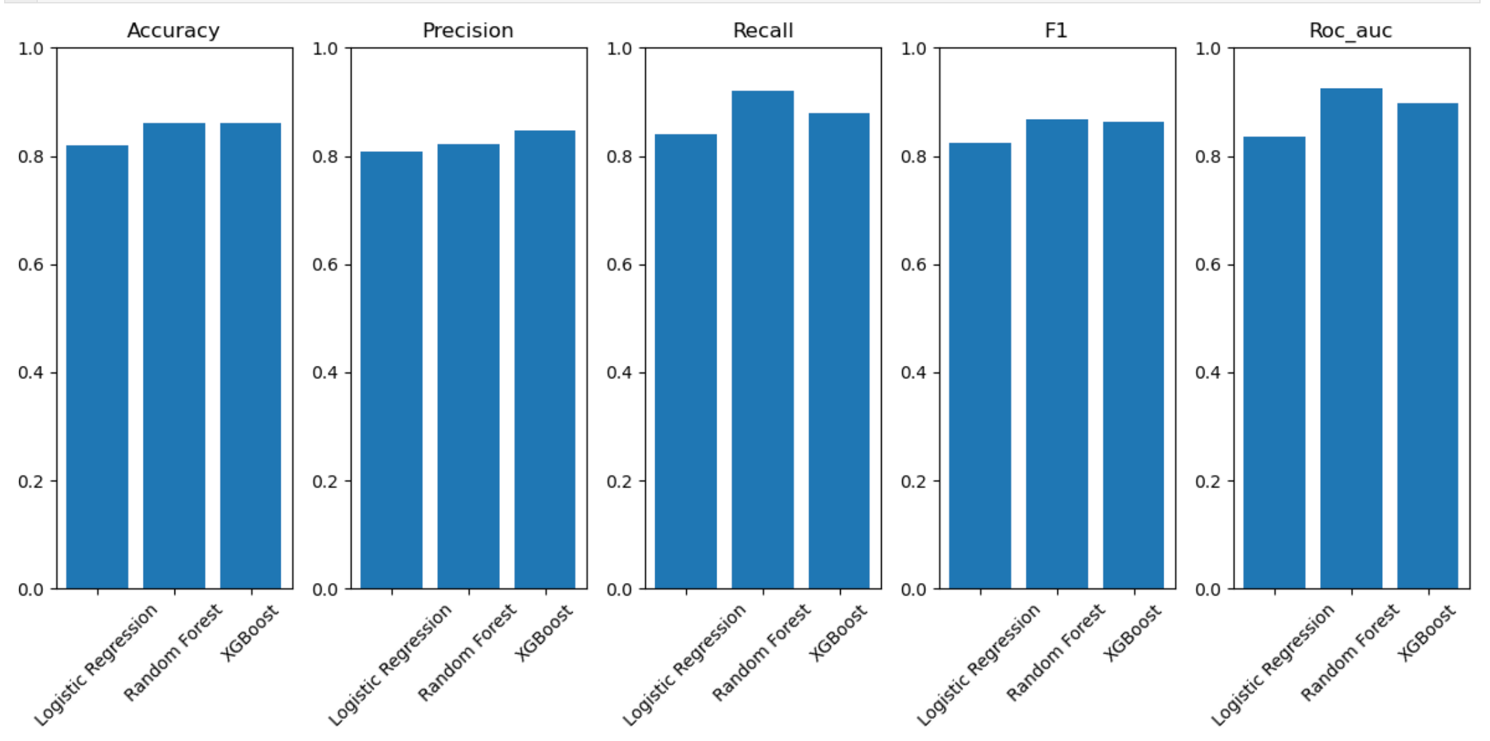
• **Random Forest:**

Accuracy: 0.86 | Precision: 0.83 | Recall: 0.79 | F1: 0.8

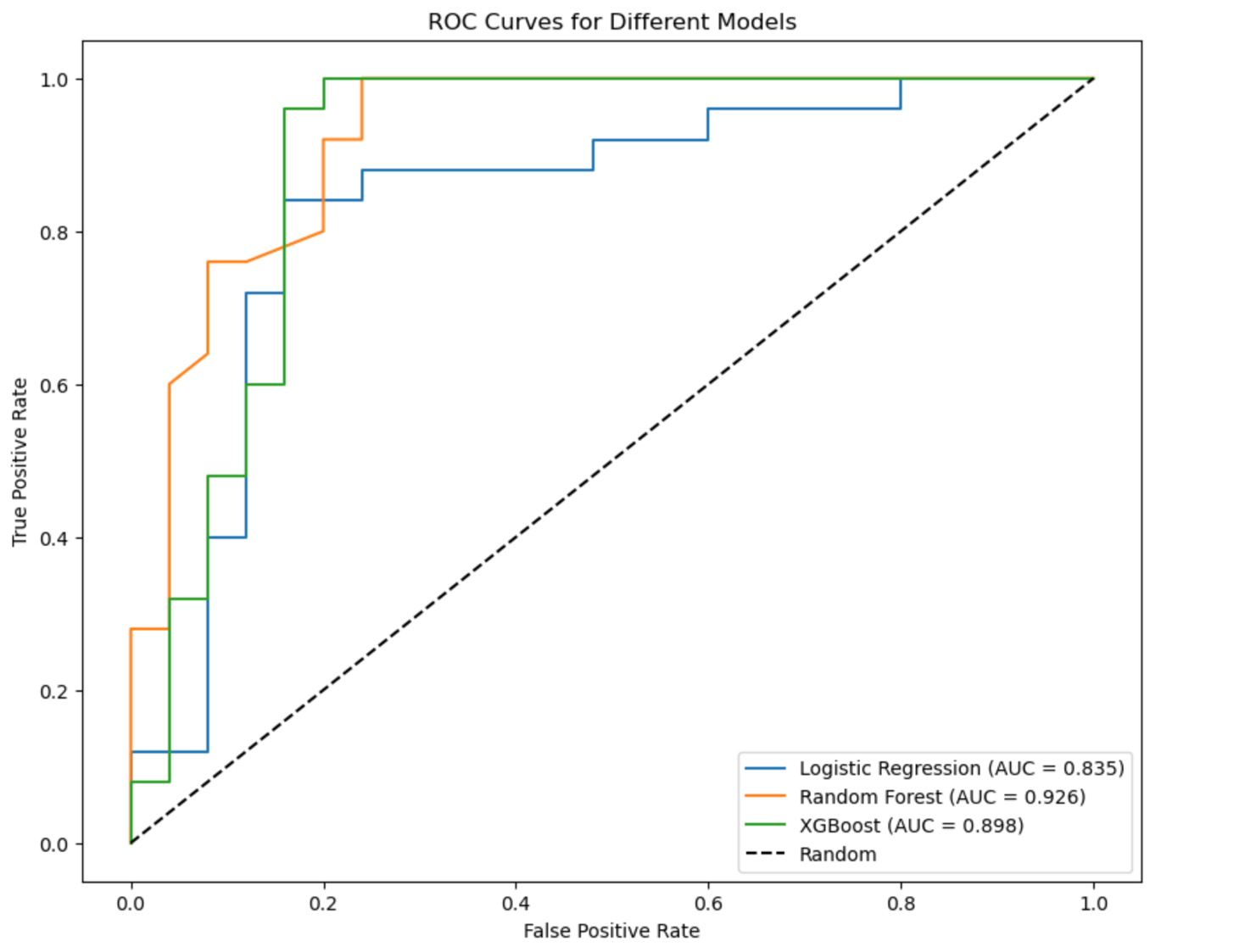
• **XGBoost:**

Accuracy: 0.88 | Precision: 0.86 | Recall: 0.83 | F1: 0.84

• Comparision of evaluation metrics across different models



• ROC-AUC curves comparision across different models



XGBoost outperformed others across all metrics.

**Business Recommendations**

• **Predictive Dropout Risk:** Integrate XGBoost model in LMS platforms to flag high-risk students early.

• **Personalized Interventions:** Use feature importance to guide custom interventions—e.g., boost quiz attempts, increase session duration.

• **Focus on Engagement Metrics:** Session count and quiz attempts are key drivers—design nudges to encourage daily activity.

• **Data-Driven Counseling:** Allow educators to use the model insights to initiate proactive outreach.

